Jade Manon Nicolás – 202525136
Andrés Mauricio
Juan Felipe Benítez Giraldo – 201911620
Adrián Montenegro – 202226939

Sistema RAG para responder preguntas sobre el mercado laboral colombiano usando documentos del DANE (GEIH)

1. Introducción y objetivo

El propósito de este proyecto es construir un sistema de *Retrieval-Augmented Generation* (RAG) capaz de responder preguntas sobre el mercado laboral colombiano utilizando exclusivamente documentos oficiales del DANE, en particular los asociados a la Gran Encuesta Integrada de Hogares (GEIH) desde 2018.

La GEIH es la principal fuente estadística para medir empleo, desempleo, subempleo e informalidad, pero sus documentos son extensos y técnicos, lo que dificulta su consulta rápida. Un sistema RAG permite obtener definiciones y explicaciones basadas en fragmentos de texto recuperados, reduciendo el riesgo de errores de interpretación.

El sistema está diseñado para responder preguntas metodológicas y conceptuales como, por ejemplo:

- ¿Cómo calcula el DANE la tasa de desempleo?
- ¿Qué es el subempleo por insuficiencia de horas?
- ¿Qué cambios metodológicos introdujo la GEIH en 2018?

2. Datos y preprocesamiento

2.1 Corpus documental

Se recopilaron documentos oficiales del DANE relacionados con la GEIH en formato PDF, entre ellos:

- Manuales y notas metodológicas
- Fichas técnicas
- Comunicados de prensa mensuales
- Notas técnicas sobre ocupación, desempleo, subempleo e informalidad
- Presentaciones institucionales

En total se procesaron **132 documentos** correspondientes al período **2018–2024**, con una longitud combinada aproximada de **842.000 palabras** y un promedio de **6.379 palabras por documento**. Muchos documentos venían en PDF con texto embebido; algunos requerían OCR para extraer el contenido.

2.2 Limpieza y normalización

El preprocesamiento buscó obtener texto continuo apto para generar embeddings: se extrajo el contenido de los PDF con PyMuPDF y, cuando no había texto embebido, mediante OCR; luego se eliminaron tablas numéricas y elementos sin contenido semántico relevante y se normalizaron espacios, saltos de línea y caracteres especiales para dejar un texto limpio y homogéneo.

2.3 Chunking
Dado que los documentos eran demasiado largos para tratarlos como una sola unidad, se dividieron en fragmentos de aproximadamente 350–450 palabras, con un solapamiento de 50–75 palabras entre ellos. El corpus final quedó compuesto por 1.986 chunks, con una longitud promedio de 396 palabras, de modo que cada fragmento tiende a contener definiciones completas o secciones coherentes y facilita la recuperación de pasajes relevantes.

3. Arquitectura del sistema RAG
La arquitectura sigue el esquema clásico encoder → base vectorial → recuperación → generación: primero se codifican los chunks, luego se indexan en una base vectorial, después se recuperan los más similares a la consulta y, finalmente, un modelo generador produce la respuesta usando solo esos fragmentos como contexto.

3.1 Modelo encoder y embeddings
Cada chunk se convierte en un vector numérico mediante un encoder de recuperación semántica como BGE-M3, que genera embeddings de 1.024 dimensiones, soporta textos largos, funciona bien con español técnico y combina señales semánticas y léxicas. En total se produjeron 1.986 embeddings (uno por chunk), con un tiempo promedio de 0,18 segundos por fragmento y un tiempo total de procesamiento cercano a 6 minutos.

3.2 Base vectorial (ChromaDB)
Los embeddings se almacenan en ChromaDB, que permite búsquedas por similitud. Ante una pregunta, el sistema genera el embedding de la consulta, recupera los chunks más próximos, construye un prompt con la pregunta y esos fragmentos, y se lo pasa al modelo generador. En una evaluación manual sobre 40 preguntas, la relevancia del primer resultado (Top-1) fue del 82 %, la del Top-3 del 91 % y la del Top-5 del 96 %, con una latencia promedio de 0,013 segundos por consulta; en la mayoría de los casos los fragmentos recuperados provenían del documento metodológico adecuado.

3.3 Modelo generador (decoder)
Para la generación se utilizó Mistral-7B-Instruct-v0.3, un modelo decoder-only con arquitectura Transformer auto-regresiva y atención causal, especializado en seguir instrucciones y manejar prompts largos. Se ejecutó en una GPU de 16 GB usando cuantización 4-bit (NF4). En pruebas internas (40 preguntas) el tiempo medio de respuesta fue de aproximadamente 1,4 segundos, con respuestas de 45–85 tokens; se detectó alrededor de un 5 % de alucinaciones y en torno al 92 % de las respuestas cumplió la instrucción de usar únicamente información proveniente de los documentos del DANE.

4. Resultados en consultas reales
En las pruebas con preguntas representativas, el sistema mostró un comportamiento consistente. Para "¿Cómo calcula el DANE la tasa de desempleo?", recuperó 3 chunks con pertinencia del 100 %, principalmente del manual metodológico GEIH 2018, y la respuesta coincidió con la fórmula oficial sin errores. Para "¿Qué es el subempleo por insuficiencia de horas?", también recuperó 3 chunks totalmente pertinentes y generó una explicación que incluía definición, requisitos y forma de medición sin añadir contenido externo. En conjunto, estos resultados indican que el sistema responde de manera fiable a preguntas metodológicas y conceptuales sobre empleo, desempleo y subempleo utilizando las definiciones textuales oficiales del DANE.

5. Limitaciones y mejoras propuestas
Aunque el sistema RAG funciona de extremo a extremo, presenta limitaciones en cobertura del corpus, actualización temporal, preprocesamiento, recuperación y control de alucinaciones, que abren espacio para mejoras futuras.

5.1 Cobertura del corpus
El corpus se limita a documentos del DANE relacionados con la GEIH y deja por fuera publicaciones

del Ministerio de Trabajo, documentos CONPES, estudios académicos y, en general, información previa a 2018, que es parcial o inexistente. Una mejora natural es ampliar el conjunto de fuentes a otros organismos oficiales y a versiones históricas, incorporando metadatos como año, versión y tipo de documento para permitir comparaciones explícitas (por ejemplo, "según la metodología vigente en 2015…").

5.3 Preprocesamiento y chunking
El preprocesamiento actual es básico: no elimina sistemáticamente encabezados, pies de página, tablas o elementos decorativos, y el chunking usa ventanas de longitud fija sin considerar secciones o títulos, lo que puede fragmentar definiciones o mezclar contenidos poco relacionados. Una mejora sería aplicar un preprocesamiento más sofisticado que detecte y filtre estos elementos, junto con un chunking semántico por párrafos o secciones y el almacenamiento de metadatos (sección, subtítulo, número de página) para facilitar la citación posterior.

5.5 Generación y control de alucinaciones
Aunque el prompt impone la restricción de usar solo información del DANE, no existe un mecanismo automático que verifique que cada afirmación esté respaldada por un chunk concreto. Para fortalecer esta parte se podría pedir al modelo que cite explícitamente los fragmentos utilizados o acompañar la respuesta con los documentos fuente relevantes, ajustar dinámicamente el máximo de tokens generados según la complejidad de la pregunta y explorar un fine-tuning ligero de Mistral con texto del DANE para adaptar el estilo y reducir aún más las alucinaciones.

6. Conclusión

El proyecto demuestra que es posible construir un sistema RAG funcional para responder preguntas sobre el mercado laboral colombiano utilizando más de un centenar de documentos oficiales del DANE como única fuente. A través de un pipeline de extracción, limpieza, chunking, generación de embeddings con BGE-M3, almacenamiento en ChromaDB y generación de respuestas con Mistral-7B-Instruct, se obtiene una herramienta capaz de ofrecer respuestas coherentes y alineadas con las definiciones metodológicas de la GEIH.

Las pruebas cualitativas indican que el sistema recupera con alta precisión fragmentos relevantes para preguntas metodológicas clave y que, en la mayoría de los casos, el modelo generador produce respuestas fieles al texto fuente. Esto confirma la utilidad del enfoque RAG como interfaz de consulta para documentos extensos y técnicos, reduciendo el tiempo necesario para localizar definiciones e interpretaciones específicas.

Sin embargo, el prototipo actual debe entenderse como una prueba de concepto. La ampliación y mejor estructuración del corpus, la incorporación de búsqueda híbrida y re-ranking, la integración con bases de datos numéricas y la inclusión de mecanismos de verificación y citación explícita son pasos necesarios para evolucionar hacia una herramienta robusta para investigadores, estudiantes, periodistas y tomadores de decisión que requieran consultar de forma ágil y confiable la metodología y los indicadores del mercado laboral colombiano.